DOCUMENT RESUME

ED 463 326 TM 033 758

AUTHOR Schwarz, Richard; Rich, Changhua; Arenson, Ethan; Podrabsky, Tracy; Cook, Gary
TITLE An Analysis of Differential Item Functioning Based on Calculator Type.
PUB DATE 2002-04-00
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Algebra; *Calculators; *High School Students; High Schools; *Item Bias; *Mathematics Tests; *Scores; State Programs; Testing Problems; Testing Programs

ABSTRACT

The effect of calculator type on student performance on a mathematics examination was studied. Differential item functioning (DIF) methodology was applied to examine group differences (calculator use) on item performance while conditioning on the relevant ability. Other survey questions were developed to ask students the extent to which they used a calculator, the perceived usefulness of the calculator, and how often it was used in the classroom. In addition, content experts were asked to identify whether an item was sensitive to calculator usage, the potential for being useful or distracting, and how the type of calculator would affect student performance. Student test data was obtained from the Tennessee Gateway assessment end-of-course test in Algebra 1. Six forms of the test were spiraled with approximately 7,000 students taking each form. No evidence of pervasive uniform difference in DIF resulting from calculator use was detected in this data. The type of questions in this assessment tended not to be sensitive to differential calculator use and did not impact test performance significantly. DIF was also not evident for students who used calculators versus those who did not. Calculator type, use, and familiarity were associated with differences in the univariate comparison of test scores. For example, students who responded that they used a graphing calculator performed higher than the other groups. Classroom practice and experience with calculators appeared to vary widely. Results from the expert judgment indicate that items with more egregious sensitivity to calculator use could be identified, although experts tended to over-identify items as being sensitive to calculator use. One appendix contains survey items related to calculator use, and the other describes the Linn-Harnisch DIF procedure. (Contains 3 figures, 4 tables, and 15 references.) (SLD)

# An Analysis of Differential Item Functioning Based on Calculator Type

By

Richard Schwarz

Changhua Rich

Ethan Arenson

Tracy Podrabsky

CTB/McGraw-Hill

Gary Cook

Wisconsin Dept. of Public Instruction





Presented at the April 2002 Meeting of NCME, New Orleans

Math calculators are commonplace in society. In education, they have been included into national content standards in mathematics (National Council of Teachers of Mathematics, 2000) and are routinely incorporated into instructional practice. Presently calculators can be grouped into four categories: four-function, scientific, graphing, and graphing with algebraic capability such as solving algebraic equations or factoring polynomials. Many assessments such as the National Assessment of Educational Progress permit calculators, but limit usage to a standard model. The usual procedure is to issue a standard calculator to all students. The use of a calculator is oftentimes also limited to a section of the test that is not calculator sensitive (i.e., calculators are deemed not to affect test performance). In concert with classroom practice, some assessments programs are now allowing greater latitude in the type of calculator used. The view is that students should not have to suddenly abandon the calculator that they have been using routinely in the classroom on an assessment. However, test standardization could be compromised or the perception of inequity can arise within the context of high-stakes assessments when students are allowed to choose their own calculators. Students could legally challenge the test results by simply stating that they were at a significant disadvantage due to the use of a calculator with less functional capacity. They could also claim inadequate opportunity to learn evidenced by a lack of instructional exposure with a certain type of calculator prior to the examination. As an aspect of test fairness, group differences should be minimized in the test design such that all students have an ample opportunity to demonstrate their ability to perform (Cole & Zieky, 2001). This paper examines group differences relating to calculator usage on a high-stakes assessment.

Previous studies on calculator effects in test situations have shown mixed results. Not surprisingly, Cohen and Kim (1992) found that for students using calculators computational types of items were easier than ones emphasizing conceptual understanding. However, items that emphasized concepts were typically more difficult when students attempted to use calculators for answering those items. Bridgeman, Harvey, and Braswell (1995) found that students who routinely used calculators in their classes typically scored higher on the SAT mathematics subtests than those who did not use them in their classrooms on a regular basis. Runde (1997) found a similar result when he compared exam scores of college algebra students who used a graphing calculator against those who did not. Using an experimental design (n=50), Hansen, Brown, Levine, and Garcia (2001) found that calculator type did not effect performance on NAEP problem sets. Scheuneman, Camara, Cascallar, Wendler and Lawerence (2002) evaluated the effect of calculator usage on SAT I performance finding that effect was small but still detectable.

One of the primary questions is to determine the effect of calculator type on performance. Given a specific type of calculator, is differential performance on certain types of items evident? For instance, do students who have calculators with graphing capability perform better than expected on items that require a function to be plotted? If performance is affected by calculator type, what are the characteristics of these math items? Are some demographic groups more affected than others by type of calculator used? How do students who elect not to use a calculator in the assessment perform?

To more definitively answer these questions, this paper applies differential item functioning (DIF) methodology. DIF methodology examines group differences (i.e., calculator usage) on item performance while conditioning on the relevant ability. That is, do students perform significantly better or worse than expected on an item given calculator type and level of ability? DIF methodology also has the advantage of providing statistical tests. Within this DIF framework, students using a particular type of calculator (e.g., four-function, scientific) can be defined as the focal group and the focus of the analysis. The focal group can then be compared against the performance of other groups. For gender or ethnic DIF analyses, there is ample precedent for defining a reference group. In performing such a DIF analysis on calculator usage, some thought has to be given to how the reference group is defined. Such a decision is somewhat arbitrary, the group with the greatest frequency could be defined as the reference group or pairwise comparisons of DIF for every group could be performed. The actuarial approach was adopted here that defined greatest frequency as the reference group and all other groups combined. Scheuneman et. al. employed a similar approach that used scientific-calculator as the reference group.

Other survey questions asked students the extent to which they utilized a calculator on the exam, the perceived usefulness of their calculator, and how often it was used in their classroom. These questions are examined in the context of a high school assessment program in mathematics with a graduation requirement.

A survey was developed in which content experts were asked to identify whether an item is sensitive to a calculator usage, the potential for being useful or distractive and which type of calculator would affect student performance as indicated in the previous question. The purpose of this survey was to determine if experts could identify items for calculator sensitivity and potential DIF a priori. This is important when developing item specifications that further support the intended interpretations and preventing items from being eliminated from the pool unnecessarily.

**Procedure**

*Instruments.* Student test data was obtained from the Tennessee Gateway assessment. The Gateway assessment is given as an end-of-course test in Algebra I, Biology, and English. A passing score on each end-of-course test is a Tennessee high school graduation requirement. Test data were obtained from a study in which 6 intact forms of the test were calibrated and equated in the spring of 2001. Each Algebra test consists of 55 selected-response items with approximately 7,000 students taking a given form. There were 330 unique items contained on the 6 test forms. The Algebra test has a mean of 500 and a standard deviation of 50. These tests did not demonstrate any evidence of students failing to complete the test (i.e., speededness). Traditional DIF analyses for gender and ethnicity were conducted with no items being flagged.

After the completion of the test, students were asked to respond to an opportunity-to-learn survey (see Appendix A). A number of questions embedded in the survey pertained to their calculator usage on the test and experience. Responses to this survey regarding type of calculator used by students were utilized as grouping variables in the DIF analyses. A judgmental review by content experts in mathematics was conducted using another survey (see Appendix A).

Method

*Scaling.* Selected-response items were scaled using the three-parameter logistic model (Lord & Novick, 1968; Lord, 1980) in which the probability that a student with ability $\theta$ responded correctly to item $i$ is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}$$

where $a_i$ denotes the item discrimination, $b_i$ the item difficulty, and $c_i$ the lower asymptote corresponding to the probability of a correct response by a very low-scoring student. The three-parameter model was estimated using marginal maximum likelihood procedures (Bock & Aitkin, 1981) via the IRT scaling program PARDUX (Burket, 1991).

*DIF Analysis.* Two DIF statistics, the Linn-Harnisch procedure (1981) and the more familiar non-parametric Standardized Mean Difference (Zwick, Donaghue, and Grima, 1993) were used for assessing DIF. Derivations for both procedures are given in Appendix B. The Linn-Harnish procedure was used primarily since it gives results based on the operational IRT score metric. Standardized Mean Difference (SMD) was used primarily as a cross-validation that utilized slightly different criteria to define the reference group. The responses to the survey questions were used as a grouping variable for the DIF analysis. The students who identified themselves as using a particular type of calculator (e.g., scientific calculator, no calculator) were the focus of the DIF analysis. In the Linn-Harnisch analyses, the observed proportion correct of a subgroup was compared with the expected proportion correct estimated using the entire calibration sample. In SMD analyses, each subgroup was compared with the subgroup that used scientific calculator. We chose the scientific calculator subgroup as the reference group because this group consists of the "majority" of the students in this test administration.

For the Linn-Harnisch procedure, an item is flagged for DIF (against or favoring a subgroup) when the observed minus the expected mean proportion correct is greater or equal than the absolute value of 0.10 and the corresponding Z value is greater and equal than the absolute values of 2.58. The Linn-Harnisch DIF analysis was conducted for each of the four types of calculator and no calculator subgroups. The SMD index expresses results in an item-level score metric. The mean SMD of -.21 indicates that on the average, the difference in mean item score (focal – reference) is more than 2/10 of a score point (Zwick et al, 1993). SMD with an absolute value of .20 and larger was flagged for DIF, that is, the items demonstrating 1/5 of a score point difference between focal and reference group comparison. The purpose for this criterion is to identify item formats with substantial sensitivity to calculator usage. The expectation was that varying items would be flagged using these differing criteria. From an exploratory approach, there was particular interest in items that were flagged by both statistics.

In order to obtain a judgmental analysis of the sensitivity of items to calculator usage, a panel of six professional content area experts in mathematics were asked to make a judgment on each item (see Appendix A). Results from this expert review were compared with the DIF analyses.

**Results**

*Sample.* The Gateway assessment was given to a census sample of Tennessee students enrolled in Algebra in the spring of 2001. Table 1 summarizes the descriptive statistics for Gateway. Six forms of the test were spiraled with approximately 7,000 students taking each one.

Table 1.

Gateway sample statistics

| Characteristic | Sample Size |
|---|---|
| Total Sample | 42010 |
| Gender | |
| Males | 21367 |
| Females | 20150 |
| Omitted | 493 |
| Ethnicity | |
| African-American | 9911 |
| White | 29119 |
| (Other & 0mitted) | 2980 |

*Descriptive Statistics.* A student Opportunity-To-Learn (OTL) survey was given following the assessment in which questions relating to calculator usage were collected. Figure number 1 gives the conditional mean scale scores based on these survey responses. The upper left panel of Figure 1, Question 1 shows that students using a graphing calculator scored appreciably higher than the other categories followed by students using a scientific calculator. Most students used scientific calculators. Interestingly, students with graphing calculators with computer algebraic system (CAS, see http://education.ti.com/product/tech/89/down/89tips-02.html) scored lower than students using scientific or graphing types. Students using four-function calculator ranked the 4th place out of 5 and students using no calculator scored the lowest. The highest frequency of students used a calculator to answer 5-10 questions (27%) followed by 11-20 questions (19%) with little appreciable difference in performance among most categories. Most students (44%) responded that calculators made the test easier followed by approximately 8,500 (20%) reporting that it made no difference. Students who responded that calculators made the test easier also scored higher than the other groups. Based on the perceived utility of using a calculator, whites were almost twice as likely to say that calculators made the test easier than African-Americans (50% versus 29%). Most students appear to use calculators with some regularity in the classroom. However, a sizable number of students (14%) responded that they were not allowed to use calculators. The majority of students (used their calculators on a daily basis for classwork or homework. However, a small number were precluded from using calculators on these assignments (5%). Whites also had a much greater tendency (87%) to respond that calculators were used with regularity either in their classroom or on homework compared with 78 percent of African-American students. Most students were allowed to use calculators on class quizzes and tests. However, a sizable number (~6000, 14%) were not allowed to use them in these circumstances. Figure 2 shows the conditional probability for ordered categories of calculator use. Whites were much more likely to use calculators on a daily basis than African-Americans.

Figure 1. Mean scale scores and sample size for a given calculator survey responses.

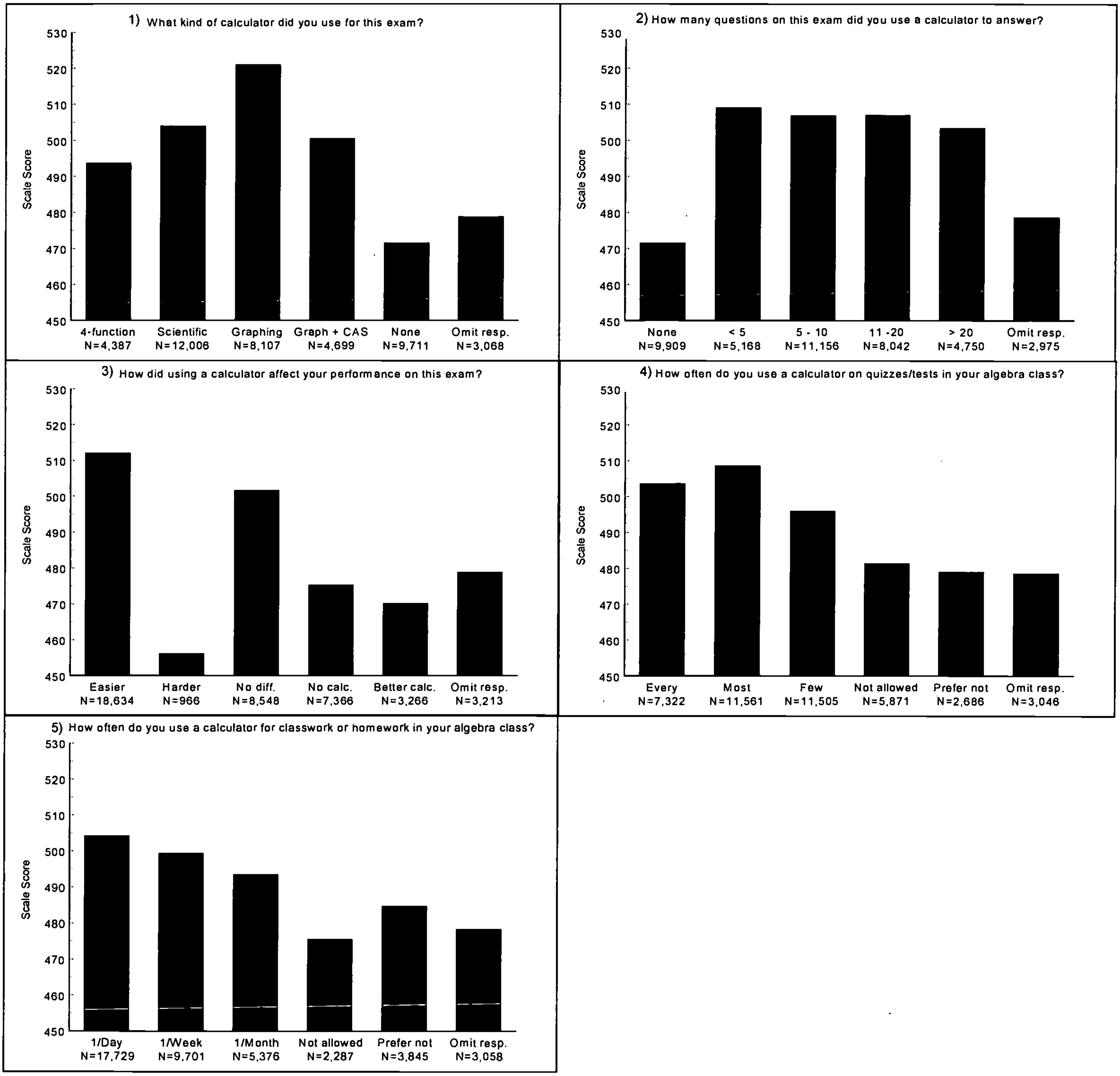

Figure 2.

Conditional probability of calculator use by ethnicity

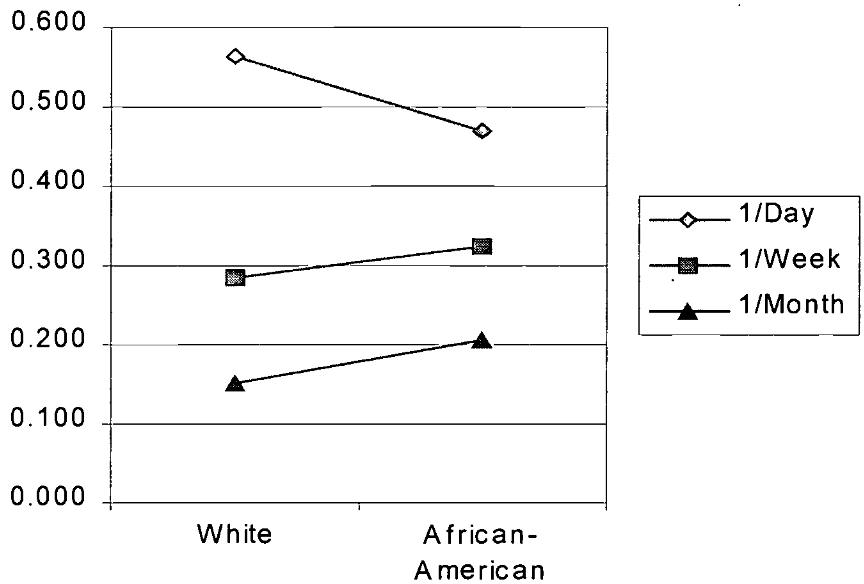


Table 2 shows calculator usage by various demographic groups. Differences for gender were the largest for scientific and the no calculator categories. African-American students compared with whites had had a lower percentage using scientific or graphing calculators and corresponding higher percentage that responded that no calculator was used. The percentage of African-American students who used graphing calculator with algebraic manipulation is comparable to the overall percentage.

Table 2.

Percentage of students by gender and ethnicity by calculator type

| Calculator Usage | All students | Female | Male | White | African-American | Other |
|---|---|---|---|---|---|---|
| 4-function | 11 | 11 | 10 | 11 | 10 | 9 |
| Scientific | 29 | 31 | 26 | 31 | 22 | 27 |
| Graphing | 19 | 19 | 20 | 21 | 14 | 17 |
| Graphing With CAS | 11 | 12 | 11 | 11 | 12 | 12 |
| No calculator | 23 | 20 | 26 | 19 | 34 | 24 |
| Omitted | 8 | 7 | 8 | 7 | 8 | 11 |

Note: CAS = computer algebraic system

*DIF results.* Since this study was a forms calibration study, several studies had occurred previously in order to iteratively purify these test forms of both gender and ethnic DIF. As a result, no items were identified as demonstrating gender or ethnic DIF using the Linn-Harnish procedure. DIF resulting from gender or ethnicity was not viewed as a possible confounding factor for this analysis.

The results from the calculator DIF analysis are shown in Table 3 for items demonstrating large amount of DIF by Linn-Harnish and/or by the SMD procedure across 6 parallel forms of the Algebra test. In general, little differential item functioning by calculator type was detected. Two items were found in favor of the students using graphing calculator with algebraic systems for both the SMD and Linn-Harnish procedures. Both DIF procedures give a mixed result on the students who did not use a calculator. One item showed negative DIF and one item showed positive DIF for the no calculator subgroup. Note that the number of students in the subgroup of "no calculator" is the second largest subgroup next to the scientific calculator group and on average students with no calculator scored appreciably lower (see Figure 1, Question 1) on the test.

It is understandable that lower ability students with no calculator might be at a disadvantage when asked to compute a mean. It is also not surprising that taking the test without a calculator was an advantage when a remainder is considered in a word problem. The graphing calculator with CAS group had comparatively lower test performance (see Figure 1, Question 1). For this group, the graphing capability improved their performance on items 48 and 54.

Table 3.

Items demonstrating calculator DIF

| Item Classification | Standardized Mean Difference | | Linn-Harnish | |
|---|---|---|---|---|
| | Graphing with CAS | No calculator | Graphing with CAS | No calculator |
| A. Item 5 – Determine the mean of a given set of real-world data (no more than five two-digit numbers) | | Against (-.196) | | Against (Z=-11.4) |
| B. Item 17 – Select a reasonable solution for a real-world division problem in which the remainder must be considered | | In favor (.169) | | In favor (Z=8.7) |
| C. Item 48 -Select the graph that represents a given linear function expressed in slope-intercept form | In favor (.188) | | In favor (Z=8.0) | |
| D. item 54 – Same classification as C (Item 48) | In favor (.181) | | In favor (Z=7.2) | |

Note: For SMD the reference group was scientific; for Linn-Harnish the reference group was all others; CAS denotes calculator with computer algebra system.

Using the results from Table 3, an examination of the operating characteristics of these items is given in Figure 3 using scientific calculator as the reference group. The expected probability (via the IRT model) of getting the item correct for a given ability level is plotted for the items flagged for calculator DIF. Test items 5 and 17 (no calculator used versus scientific) demonstrate greater differences at the bottom part of the distribution in the probability of getting the item correct. The bubble plot below shows that few students are in this region. The upper parts of the ability distribution for these two items show good agreement. By contrast, the comparison of scientific versus calculators with CAS showed differences throughout the ability distribution. Students with scientific calculators had to have higher scale scores in order to have the same probability of getting the item correct compared with the calculator with CAS group. Note that the CAS group had comparatively lower scores compared with the scientific or graphing groups. Therefore, DIF emerged for this type of calculator but not the graphing group that had higher scores.

In addition to the DIF analysis, a survey on the impact of calculator use was conducted using six content experts in mathematics. These experts were asked to judge the calculator sensitivity of each item on a designated test form by answering three survey questions for each of the 55 items. The purpose of the survey was to determine if judges could identify the items flagged for DIF. The results of this analysis are shown in Table 4. All 6 experts responded that the use of four-function calculator would be helpful on item 5, which was flagged for DIF against the no calculator subgroup. Opinions were split on items 17 and 48. Item 17, which was flagged for DIF in favor of no calculator subgroup, two experts chose helpful and 2 experts chose "distractive". For the third question, two experts chose distractive to the second question and marked all 4 types of calculators as distractive. On item 48, the experts who responded "helpful" indicated that both types of graphing calculators would be useful, which is basically consistent with the DIF analysis.

Figure 3. Item characteristic curves for items flagged for calculator DIF

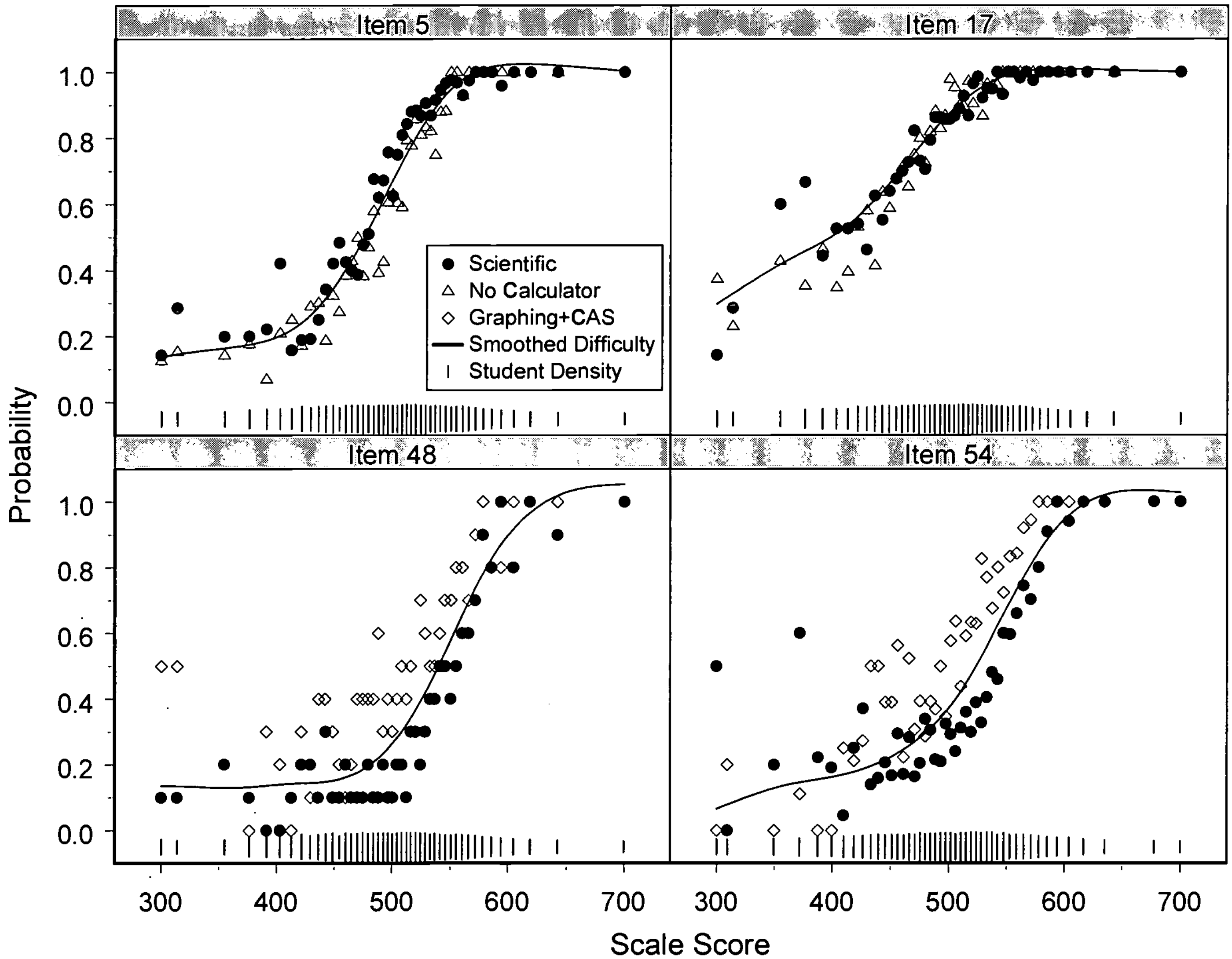

Table 4.

Judges responses for calculator sensitivity for items flagged for DIF

| Question 1: Does calculator affect student performance? | Frequency | |
|---|---|---|
| | Yes | No |
| Item 5 | 6 | 0 |
| Item 17 | 4 | 2 |
| Item 48 | 2 | 4 |
| Question 2: Helpful or Distractive? | Helpful | Distractive |
| Item 5 | 6 | 0 |
| Item 17 | 2 | 2 |
| Item 48 | 2 | 0 |

Note: Item 54 was on an alternate form that was not judged; n =6 judges.

The experts were also in agreement on other items not flagged by DIF statistics in which they agreed 83 percent of time or greater. The survey analysis shows that both DIF analysis and the expert's review of items for calculator DIF are important. However, if we were to depend on the expert review only, nearly one third of items would be considered as showing some degree of advantage for students with various types of calculators.

## Conclusion

The examination of DIF can be viewed as an aspect of test fairness where group differences should be minimized in the test design such that all students have an ample opportunity to demonstrate their ability to perform. On mathematics examinations in which calculators are not standardized, a differential impact on performance might exist. No evidence of pervasive uniform differences in DIF due to differential calculator usage was detected for this data. The type of algebra questions contained is this assessment tended not to be sensitive to differential calculator usage and did not significantly impact test performance. Only a few items were flagged for DIF across multiple forms using two slightly varying definitions for the reference group. DIF also was not evident for students who used calculators versus those that did not. DIF with respect to calculator remains an empirical question that is subject to the particular type of items on the test. Calculators can either enhance validity or be a source of construct irrelevant variance if calculator sensitivity is not taken into account before constructing the test blueprint (Bridgeman, Harvey & Braswell, 1995). For instance, it is common practice in achievement tests used in the elementary grades to designate sections of the test with items that are conceptual in nature in which calculators are permitted. Other sections of the test, which are computational in nature, exclude their use. Items that are conceptual or ones that do not require much computation or graphing functions will tend not to be calculator sensitive. It is reasonable to suppose that items that involve computation and are easy, low ability students might benefit from the use of a calculator. This was the case in Gateway when a question required the mean to be computed. Similarly, lower ability students benefited when calculators with advanced functions were used to solve items that required a function to be identified. In our judgement, none of the item types flagged in this paper need be eliminated from the item selection. For two items (5 & 17), group differences were mitigated by DIF cancellation (Shealy & Stout, 1993 ) when "no calculator" was the focal group. Secondly, performance was relatively uniform (Question 2) despite how many questions were answered using a calculator. This indicates that the test is not calculator sensitive.

Calculator type, usage, and familiarity were associated with differences in the univariate comparsion of test scores. For instance, students who responded that a graphing calculator was used performed higher than the other groups. The use of a graphing calculator could indicate that higher-level mathematics courses had been taken. However, this may not be the case in this instance since all students take the Gateway examination at the conclusion of their first Algebra I class. Relatively large differences were found between various demographic groups in calculator usage and experience. These differences were also noted by Scheuneman et al. (2002) where calculator usage was investigated for students taking SAT I. The patterns of usage for that study and this one are similar since the population of students are somewhat comparable; that of high school students. Female and male students had comparable percentages with respect to type of calculator used. However, the calculator usage between African-Americans and whites differed more markedly. African-Americans were more likely not to use a calculator. Also, classroom practice and experience with respect to calculator usage appears to vary greatly. Significant equity concerns could have emerged if these Gateway items were sensitive to calculator type or familiarity.

The results from the expert judgement indicate that items with more egregious sensitivity to calculator usage could be identified. However, expert judgement tended to over-identify items as being calculator sensitive. This suggests, in concert with traditional DIF analyses, that judgemental review be performed as a first step; followed by a DIF analyses to definitely address calculator sensitivity.

DIF has been a well-studied problem and is routinely a step in the test validation process for gender and ethnicity. Schwarz (1998) suggested that DIF methodology should be applied to new contexts. The analysis of calculator DIF represents an additional area in which group differences can be examined. Survey data when used in concert with test information can be used to examine group differences of interest on the ability metric.

## References

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32*, 323–340.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443–459.

Burket, G. R. (1991). *PARDUX* [Computer program]. Unpublished.

Cohen, A.S., & Kim, S.-H. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5*, 303–320.

Cole, N.S., & Zieky, M.J. (2001). The New Faces of Fairness. *Journal of Educational Measurement, 38*, 369–382.

Hansen, K., Brown, B., Levine, R., and Garcia, T. (2001). Should standard calculators be provided in testing situations? An investigation of performance and preference differences. *Applied Measurement in Education, 15*, 59–72.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement, 18*, 109–118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (71, 179–181). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. 402 pp.

Runde, D. C. (1997). The effect of using the TI-92 on basic college algebra students' ability to solve word problems. (Available from ERIC document reproduction service no. ED409046).

Scheuneman, J.D., Camara, W.J., Cascallar, A.S.,Wendler, C., and Lawrence, I. (2002). Calculator Access, Use, and Type in Relation to Performance on the SAT I: Reasoning Test in Mathematics. *Applied Measurement in Education, 15*, 95–112.

Schwarz, R.D. (1998). Tracelines for Classification Decisions. *Applied Measurement in Education, 11*, 311-330.

Shealy, R.T., & Stout, W.F. (1993). An item response theory model for test bias and differential test functioning. In P.W. Holland and H. Wainer (Eds), Differential Item Functioning (pp. 197–239). New Jersey: Lawrence Erlbaum Associates.

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

## Appendix A

*Student survey items pertaining to calculator usage*

1. What kind of calculator did you use in this exam?

   O 4-function calculator
   O Scientific calculator
   O Graphing calculator
   O Multi-functional calculator with algebraic capabilities
   O I did not use a calculator

2. How many questions in this exam did you use calculator to answer?

   O less than 10
   O 10-20
   O 21-30
   O 31-40
   O 41-50

3. How do you feel about the use of a calculator in this exam

   O made the test easier
   O made the test more difficult
   O made no difference

4. How often do you use a calculator on tests in your math class?

   O often
   O sometimes
   O never, not allowed
   O never, I prefer not to

*Survey of calculator sensitivity by content experts*

1. Would the use of a calculator affect student performance on the item?

   O Yes
   O No
   O Not Certain

2. If yes, is calculator usage helpful or distractive?

   O Helpful
   O Distractive

3. If helpful or distractive, which type of calculator would affect student performance as indicated in question 2?

   O Four-function
   O Scientific
   O Graphing
   O Graphing with computer algebraic system (CAS)
   O No calculator

## Appendix B

Linn-Harnisch. The Linn-Harnisch procedure uses the systematic differences between the obtained and expected frequencies derived via the three-parameter model . First, the sample is divided into ten equal score categories (deciles) based upon their location on the ability score ($\theta$) scale for a given item. The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group. The proportion of people in decile *g* who are expected to answer item *i* correctly is

$$P_{ig} = \frac{1}{n_g} \sum_{j \varepsilon g} P_{ij}, \tag{1}$$

where $n_g$ is the number of examinees in decile *g*. The proportion of people expected to answer item *i* correctly (over all deciles) for a group (e.g., students who use a scientific calculator) is:

$$P_{i \cdot} = \frac{\sum_{g=1}^{10} n_g P_{ig}}{\sum_{g=1}^{10} n_g} \tag{2}$$

The corresponding observed proportion correct for examinees in a decile ($O_{ig}$) is the number of examinees in decile *g* who answered item *i* correctly divided by the number of students in the decile ($n_g$). That is,

$$O_{ig} = \frac{\sum_{j \varepsilon g} u_{ij}}{n_g}, \tag{3}$$

where $u_{ij}$ is the dichotomous score for item *i* for examinee *j*.

The corresponding formula to compute the observed proportion, over all deciles, of students answering each item correctly in the group is given by:

$$O_{i\cdot} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}, \tag{4}$$

After the values are calculated for these variables, the difference between the observed proportion correct and expected proportion correct for a particular group can be computed. The decile group difference ($D_{ig}$) for observed and expected proportion correctly answering item $i$ in decile $g$ is

$$D_{ig} = O_{ig} - P_{ig}, \tag{5}$$

and the overall group difference ($D_{j}$) between observed and expected proportion correct for item $i$ in the complete group (over all deciles) is

$$D_{i\cdot} = O_{i\cdot} - P_{i\cdot}\,. \tag{6}$$

These indices are indicators of the degree to which members of a group perform better or worse than expected on each item, based on the parameters estimated from all groups. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) scale. The decile group difference ($D_{ig}$) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference.

Items are flagged as demonstrating DIF for (+) or against (*) the specified subgroup according to the following rule: An item demonstrates DIF against a subgroup if the $D_{i.} \leq -0.10$ and $Z \leq 2.58$. DIF in favor of a subgroup is defined in the same way but with a positive difference.

Standardized Mean Difference. The Standardized Mean Difference (SMD) is an extension of the Mantel-Haenszel (MH) statistic used for calculating DIF where

$$\text{SMD} = \Sigma\, p_{Fk}\, m_{Fk} - \Sigma\, p_{Fk}\, m_{Rk}, \tag{7}$$

where

$$p_{Fk} = n_{F+k} / n_{F++} \tag{8}$$

is the proportion of focal group members who are at the kth level of the matching variable,

$$m_{Fk} = (1/n_{F+k})\,(\Sigma\, y_{t}\, n_{Rtk}) \tag{9}$$

is the mean item score for the focal group at the kth level, and

$$m_{Rk} = (1/n_{R+k})(\Sigma\, y_{t}\, n_{Rtk}) \tag{10}$$

is the analogous value for the reference group. A positive value for a SMD reflects DIF in favor of the focal group. Likewise, a negative SMD reflects DIF against the focal group.

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*

ERIC®

TM033758

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: An Analysis of Differential Item Functioning Based on Calculator Type

Author(s): Richard Schwarz, Changhua Rich, Ethan Arenson, Tracy Podrabsky, + Gary Cook

Corporate Source: CTB/McGraw-Hill

Publication Date: April 2002

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

| The sample sticker shown below will be affixed to all Level 1 documents | The sample sticker shown below will be affixed to all Level 2A documents | The sample sticker shown below will be affixed to all Level 2B documents |
|---|---|---|
| PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY<br>Sample<br>TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)<br>1 | PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY<br>Sample<br>TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)<br>2A | PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY<br>Sample<br>TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)<br>2B |
| Level 1 | Level 2A | Level 2B |
| ☒ | ☐ | ☐ |
| Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy. | Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only | Check here for Level 2B release, permitting reproduction and dissemination in microfiche only |

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

*I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.*

Sign here,→ please

| | | |
|---|---|---|
| Signature: Tracy L. Podrabsky | Printed Name/Position/Title: Tracy Podrabsky / Research Associate | |
| Organization/Address: CTB/McGraw-Hill<br>20 Ryan Ranch Rd, Monterey, CA 93940 | Telephone: 831-393-7862 | FAX: 831-393-7016 |
| | E-Mail Address: tpodrabsky@ctb.com | Date: 3/25/02 |

(over)

## III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

| |
|---|
| Publisher/Distributor: |
| Address: |
| Price: |

## IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

| |
|---|
| Name: |
| Address: |

## V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

**University of Maryland**
**ERIC Clearinghouse on Assessment and Evaluation**
**1129 Shriver Laboratory**
**College Park, MD 20742**
**Attn: Acquisitions**

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

**ERIC Processing and Reference Facility**
**4483-A Forbes Boulevard**
**Lanham, Maryland 20706**

**Telephone: 301-552-4200**
**Toll Free: 800-799-3742**
**FAX: 301-552-4700**
**e-mail: ericfac@inet.ed.gov**
**WWW: http://ericfac.piccard.csc.com**

EFF-088 (Rev. 2/2000)